



# Perception of voicing in the absence of native voicing experience

Rikke Louise Bundgaard-Nielsen<sup>1,2</sup>, Brett Baker<sup>3</sup>

<sup>1</sup> MARCS Institute, University of Western Sydney, Australia

<sup>2</sup> La Trobe University, Australia

<sup>3</sup> Melbourne University, Australia

rikkelou@gmail.com, bjbaker@unimelb.edu.au

## Abstract

50 years of speech perception research provide a rich literature on cross- and second language (L2) perception of stop consonant contrasts such as /p b/, /t d/, and /k g/ which differ systematically in the relative timing of oral stop release and the onset of vocal fold vibration (voice onset time: VOT). This research has focused primarily on two observations: 1) that nonnative listeners automatically use their native VOT contrast boundary in an unfamiliar language, irrespective of whether this language shares the boundary or not, and 2) that even highly proficient L2 language users often perceive L2 VOT-based contrasts in a way that is consistent with their L1, even after decades of L2 acquisition. No work has, hitherto, examined VOT-based contrast discrimination by L1 speakers of languages without any VOT-based stop contrast. In the following, we present two studies of speakers in such a scenario, showing that even extensive L2 experience is insufficient for L2 learners without native (L1) voicing experience to acquire such a distinction.

**Index Terms:** speech perception, phonology, language acquisition.

## 1. Introduction

50 years of speech perception research provides a rich literature on cross- and second language (L2) perception of stop consonant contrasts such as /p b/, /t d/, and /k g/. Stop pairs like these are produced with similar tongue configurations forming a complete blocking of oral airflow at the labial (/p b/), alveolar (/t d/) or velar (/k g/) place of articulation but differ systematically in the relative timing of oral stop release and the onset of vocal fold vibration (voice onset time: VOT).

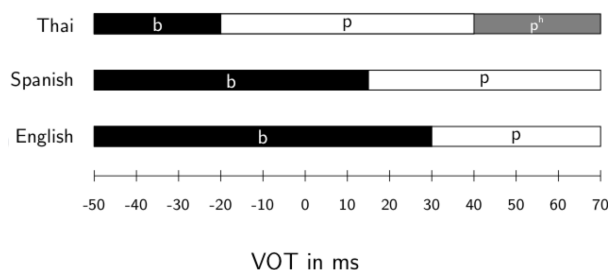


Figure 1: VOT boundaries across three languages. Adapted from [1].

Across the world's languages, VOT is typically considered to be a 3-way distinction between voiced/voiceless/voiceless aspirated ([1]; see Fig. 1). Most languages have two categories (English, Spanish, Mandarin), some have three (Thai, Korean), and a few have four categories (Hindi, Marathi). Very few languages make no use of voicing distinctions but most Australian languages, including Wubuy (see section 2), fall into this category.

Perception research focusing on voicing-based stop distinctions has made two important observations. Firstly, that nonnative listeners systematically and automatically use their L1 VOT contrast boundary in perceiving phones in an L2, even when the phonetic realisation of the contrast in the L2 differs from that of the listeners' L1, such as is the case of Spanish learners of English. Secondly, research has shown that this phenomenon of 'accented perception' often persists also for highly proficient L2 language users. Indeed, this has been shown on the *phonological* level for English learners of Thai who struggle to discriminate the three Thai stop VOT categories [1], [2], [3], and on a *phonetic* level in, for instance, the difficulty experienced by Spanish-English bilinguals whose languages differ in the VOT setting for voiced and voiceless stops (shortlag/longlag versus prevoiced/shortlag stop realisation) [4], [5].

Such evidence for L2 accented perception has been interpreted from a number of theoretical perspectives, including the Perceptual Assimilation Model (PAM: [6], [7]; PAM-L2: [8]). According to PAM/PAM-L2, L1 phonological learning shapes the way in which L2 phones are perceived. Indeed, PAM predicts that L2 phones are discriminated on the basis of their mapping into L1 phoneme categories in a number of different patterns. Those relevant for the present study are (1) **Single Category** (SC) contrasts in which two L2 phones are perceived as equally good instances of the same L1 phonemic category, and discrimination is expected to be poor, (2) **Category Goodness** (CG) contrasts in which two L2 phones are instances of the same L1 phonemic category, but one L2 phone is perceived as a 'better' fit than the other, and discrimination is predicted to be moderate to good and (3) **Two-Category** (TC) contrasts where two L2 phones are assimilated into separate L1 phoneme categories. Discrimination is predicted to be excellent.

No work has, hitherto, examined VOT-based contrast discrimination by L1 speakers of languages without voicing-based contrasts altogether, such as Wubuy, and which might provide crucial insights into the question of how flexible the speech perception system is when it comes to a dimension of speech not exploited in the native language. In the following,

10.21437/Interspeech.2015-509

we present two studies of speakers in such a scenario. Study 1 tested the discrimination of English stop consonants and Study 2 the discrimination of English fricatives and a fricative-stop contrast by two groups of listeners from the remote Aboriginal settlement of Numbulwar in Arnhem Land (NT, Australia), who differ systematically in L1 VOT experience: 8 speakers of the Indigenous language Wubuy, and 9 speakers of Roper Kriol (see Section 2). We also tested a control group of native English speakers, to ensure that the English contrasts are indeed easily discriminated by L1 listeners.

## 2. Roper Kriol and Wubuy

Roper Kriol is an English-lexified creole which developed in the drainage basin of the Roper River in the late 19th and early 20th century as a result of contact between English speakers and speakers of traditional Indigenous languages ([9], [10], [11]). It is a lingua franca throughout South-Eastern Arnhem Land and adjacent regions, and a major variety of the largest Indigenous language in Australia. There are an estimated 20,000 L1 speakers of Kriol [12] and closely related varieties such as Fitzroy Crossing Kriol [13] across Northern Australia.

According to recent work, the obstruent inventory of Roper Kriol is English-like in its stop voicing distinction: the Roper Kriol stop contrast is based on a short-lag versus long-lag VOT difference and relies on a vowel duration difference in syllable-final positions, just as in English. Kriol however differs from English in the absence of voicing-based fricative contrasts. Instead, all fricatives are voiceless in every position in the word. However, Roper Kriol is also unquestionably similar to the substrate languages in terms of the constriction durations of stops: Kriol voiced and voiceless stops differ not only in terms of VOT, but also, in a decidedly un-English-like fashion, in terms of their constriction duration, with voiceless stops having much longer duration than voiced stops. The obstruent inventory of Roper Kriol [14] is presented in Table 1a.

Wubuy (a.k.a. 'Nunggbuyu', as in [15]) is an endangered traditional Indigenous Australian language spoken in south-eastern Arnhem Land, around the southern part of Blue Mud Bay. It is the first language for adults over the age of around 55 in the community of Numbulwar, NT, as well as a first or second language for many adults in neighbouring Groote Eylandt. Children are no longer acquiring Wubuy as a first language, though all children are exposed to Wubuy through the language revitalisation efforts at Numbulwar school. There are perhaps 60 fluent L1 Wubuy speakers. The obstruent inventory of Wubuy is presented in Table 1b. The phonology of Wubuy resembles the neighbouring Yolngu languages in having the rare four-way coronal place distinction among the stops /t̪, t, t̥, t̥̰/. It does not have a voicing distinction in stops, nor a fortis-lenis contrast (found in other languages in the area; [16]). Like most other Australian languages, Wubuy is also unusual cross-linguistically in that it has no fricatives (though see note on marginal /s/ in section 3.3).

Table 1a. *The obstruent inventory of Roper Kriol.*

Lab.	Dent.	Alv.	Retrofl.	Alv.- pal.	Vel.	Glott.
p b	t̪ d̪	t d	d̪		k g	
f		s		tʃ dʒ		h

Table 1b. *The obstruent inventory of Wubuy.*

Lab.	Lam.- dent.	Apic.- Alv.	Apic.- postalv.	Lam.- postalv.	Vel.
p	t̪	t	t̪	t̪	k

## 3. Method

### 3.1. Stimuli

We recorded three female speakers of Australian English in a recording studio at Melbourne University. All speakers were from the Greater Melbourne area in Victoria, Australia, and all had native English-speaking parents. All had substantial linguistic training. None reported fluency in any other language, though all had studied other languages in a school or university setting.

Each of the three speakers produced five repetitions of each of the target consonants /p b k/ in an /aCa/ (i.e. intervocalic) context for Study 1, and five repetitions of the target consonants /b v s z ʃ/ in a /##Ca/ (i.e. utterance-initial) context for Study 2. The participants were encouraged to familiarise themselves with the nonsense words prior to the recording and all dysfluent and mispronounced tokens were re-recorded. During the recording, the women were instructed to speak in a clear, comfortable voice as though they were speaking to a friend. All recordings had a 16-bit sampling depth with a sampling rate of 44.1 KHz.

All tokens were segmented by the first author and preceding vowel duration and F0-F3 (Study 1) and following vowel duration and F0-F3 (Study 1 and 2), VOT and constriction duration were extracted using a *praat* script [17]. Three tokens per target consonant per speaker (9 unique tokens) selected as stimuli for the perception studies on the basis of speaking rate and similar F0.

In order to test discrimination of a Kriol-like /p-b/ voicing distinction, i.e., one which is maintained by VOT as in English and also by constriction duration, we manipulated the duration of the English /p/ tokens selected for Study 1 to create a 'Kriol-like' /p/ (henceforth /p+̰/). The average constriction duration difference between Kriol /p/ and /b/ is approximately 60 ms, in clear lab-like speech, commensurate with the speech used in the present study [14]. The average /p-b/ CD difference in the English targets recorded for this study is approximately 10 ms, and we thus introduced 50 ms of silence, mid closure, for each intervocalic English /p/ token from Study 1, in order to create matching /p+̰/ tokens, while maintaining natural variation. Finally, each excised token was enveloped with a 20 ms ramp-in and a 10 ms ramp-out.

### 3.2. Experimental design

We presented two randomised cross-speaker categorical XAB discrimination tasks to Wubuy, Kriol and Australian English listeners (control group). Study 1 tested discrimination of English intervocalic stops /p k/, /p b/, and the Kriol-like manipulated contrast of /p+̰ b/. Study 2 tested discrimination of syllable-initial English fricatives /s ʃ/ and /s z/, and syllable initial /b v/. The tests were programmed in Psychscope X [18], with the stimuli presented over headphones from a MacBook computer. For both studies, the inter stimulus interval (ISI) was 500 ms. The response window was presented for 3 seconds, and all missed trials were replayed later, at a random time. The inter-trial interval was 1 second. Each of the total of six contrasts (/p k/, /p b/, /p+̰ b/ in Study 1, and /b v/, /s ʃ/ and

/s z/ in Study 2) were presented to the listeners in 6 unique triads, with 12 repetitions per triad, equaling 72 triads/contrast per listener. The task was explained to the participants as one in which a 'teacher' (first voice) was being imitated by a 'good student' and a 'bad student' (voices 2 & 3), and the participant had to indicate (with a key press on the keyboard) which was the 'good student' who copied the teacher correctly. All participants completed Study 1 first. This approach had been successfully used with Wubuy speakers in previous studies.

### 3.3. Participants

**Wubuy:** 8 native speakers (age approx. 25-65). Some were literate and some semi-literate in Wubuy. All spoke and read English (the medium of instruction at school) and the community language Roper Kriol to varying levels of proficiency (see section 5). Another four Wubuy speakers were tested but excluded from the analyses for the following reasons: three failed to understand the task, and one was reluctant to complete the task. All testing took place in a quiet home in Numbulwar, NT. All procedures were explained in English by the authors and in Wubuy by a native speaker, when needed. Each participant was compensated for their time and effort by a \$100 payment.

**Kriol:** 9 native speakers of Kriol (age approx. 18-50). All were literate (to some extent) in English and had some competence in reading and writing Wubuy and Kriol. Kriol is not formally taught at school in the region, and while some participants had some Kriol instruction through church activities, others were autodidact, mainly through the use of social media (texting on mobile phones, facebook etc.). The testing conditions and compensation were identical to those of the Wubuy speakers. A Kriol translator was available when needed.

**Australian English:** 13 native speakers of Australian English (M age 20; range 18-33). All were University of Melbourne undergraduates and recruited by word of mouth. Most had some competence in at least one other language acquired through formal instruction. One participant was excluded due to a history of learning disorders, another due to having Italian-speaking background: Italian VOT distinctions differ systematically from English, and Italian features long and short consonants (one of the parameters tested in our study). All testing took place at University of Melbourne. All procedures were explained to the participants by the authors or a research assistant. Each participant was compensated for their time and effort by a \$30 payment.

### 3.4. Predictions

On the basis of PAM/PAM-L2 ([6], [7], [8]), we can make a number of language-specific predictions, outlined below, and one shared prediction: **All** listeners will discriminate the (TC) control contrast /p k/ successfully. **Wubuy** listeners will perceive /p p+ b/ as instances of Wubuy /p/ and fail to discriminate them (SC contrast). Discrimination of /b v/ will be moderate as listeners will perceive /b/ as a good and /v/ as a 'less good' instance of Wubuy /b/ (CG discrimination). Discrimination of /s f/ will be moderate due to 1) experience with multiple place of articulation contrasts in the alveolar region and, 2) the occurrence of /s/ in a single, highly frequent, word (/sa!/ used to shoo away dogs), resulting in listeners perceiving /s/ as a 'good' and /f/ as a 'less good' instance of marginal Wubuy /s/ (CG contrast). Discrimination of /s z/ will be poor as both are instances of /s/ and listeners have no LI

experience with fricative voicing contrasts (SC contrast). **Kriol** speakers will perceive /p p+ b/ as instances of Kriol /p/, and /b/ as Kriol /b/ and discriminate them though /p b/ will be discriminated less successfully than /p+ b/ due to the lack of native Kriol-like duration differentiation (TC contrasts). Discrimination of /b v/ will be moderate as Kriol listeners will perceive /b/ as a good and /v/ as a 'less good' instance of Kriol /b/ (CG discrimination). Discrimination of /s f/ will be excellent as this is a native TC contrast. Finally, discrimination of /s z/ will be poor as both are instances of /s/ and Kriol speakers have no experience with fricative voicing (SC contrast). **English** listeners will successfully discriminate all native contrasts, including the enhanced Kriol-like /p+ b/ contrast.

## 4. Results

Figures 2 and 3 present the discrimination results from Study 1 (Fig. 2) and Study 2 (Fig. 3). As predicted, there appear to be large differences in the discrimination success of the three listener groups, depending on the contrast in question.

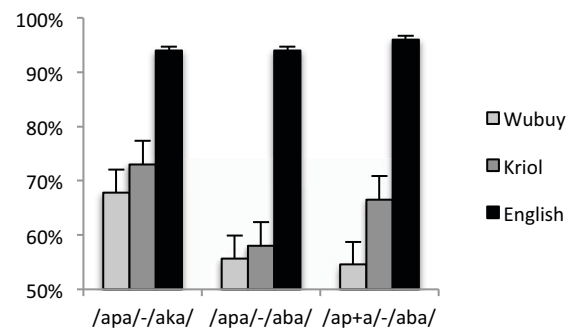


Figure 2: Mean discrimination accuracy for Wubuy, Kriol and English speakers in Study 1. Error bars indicate S.E.

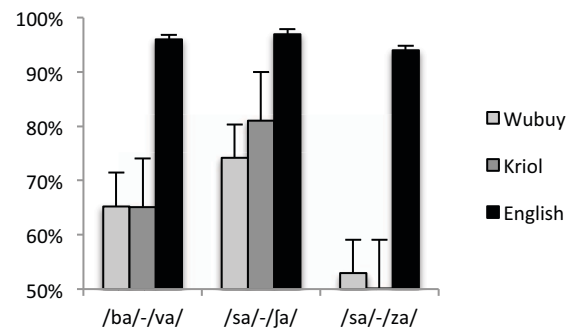


Figure 3: Mean discrimination accuracy for Wubuy, Kriol and English speakers in Study 2. Error bars indicate S.E.

### 4.1. Wubuy results

To assess whether the Wubuy listeners were able to discriminate the target contrasts in Studies 1 and 2, we first conducted a series of one-sample t-tests against chance performance. The results indicate that the Wubuy speakers are able to discriminate four of the six contrasts above chance (/p k/:  $p = .01$ ; /p b/:  $p = .025$ ; /s f/:  $p < .001$ ; and /b v/:  $p = .013$ ). The Wubuy speakers' discrimination accuracy for the Kriol-like /p+ b/ ( $p = .079$ ) and /s z/ ( $p = .204$ ) did not differ

significantly from chance performance. Two separate One-Way ANOVAs revealed a significant main effect of contrast for each of the two studies (Study 1:  $F(2,28) = 6.275, p = .006$ ; Study 2:  $F(2,28) = 12.535, p < .001$ .) Post-hoc Bonferroni-corrected comparisons confirmed that the main effect of Contrast in Study 1 was due to a significant discrimination accuracy difference between /p k/ and the other two contrasts (/p b/ and /p+ b/:  $p = .015$  for both). Post-hoc Bonferroni comparisons of the difference in contrast discrimination in Study 2 likewise confirmed that the main effect was due to the poor discrimination accuracy for /s z/ relative to /s j/ and /b v/ ( $p = .001$  for both).

The results from Study 1 are fully consistent with the PAM-based predictions above and suggest that L2 acquisition of voicing-based contrasts is extremely difficult when the learner's L1 has led him/her to consistently ignore the feature 'voicing'. The Wubuy speakers find English VOT-based labial stop contrasts very difficult to discriminate. This is also true of the Kriol-like labial stop contrast based on VOT and duration differences: unlike other languages of the area, such as nearby, related Ngandi [19], Wubuy does not implement a stop contrast based on duration or any other correlate. The results from Study 2 are also consistent with the predictions: Wubuy speakers are unable to discriminate the voicing-based fricative distinction /s z/, though they can discriminate the CG contrasts /s j/ and /b v/.

#### 4.2. Kriol results

To assess the Kriol discrimination performance, we first conducted a series of one-sample *t*-tests against chance performance which indicate that the Kriol speakers are able to discriminate all contrasts above chance level ( $p < .01$  for /p k/, /p b/, /s j/;  $p = .05$  for /p+ b/). The contrast /b v/ approached significance ( $p = .06$ ); but /s z/ was clearly not significantly different from chance ( $p = .956$ ). Two separate One-Way ANOVAs revealed a significant main effect of contrast for each of the two studies (Study 1:  $F(2,30) = 4.386, p = .021$ ; Study 2:  $F(2,27) = 16.017, p < .001$ ). Subsequent Bonferroni post hoc comparisons revealed that the main effect in Study 1 was due to English /p b/ being less accurately discriminated than the control contrast /p k/ ( $p = .018$ ). There was no significant difference in discrimination accuracy for /p+ b/ and /p b/ ( $p = .316$ ), nor in the discrimination accuracy of /p k/ and /p+ b/ ( $p = .627$ ). In Study 2, Bonferroni post hoc comparisons revealed that /s j/ differed significantly from /b v/ ( $p = .018$ ) and /s z/ ( $p < .001$ ). /b v/ also differed significantly from /s z/ ( $p = .037$ ).

The results from Study 1 suggest that Kriol speakers rely on duration as a means of distinguishing the voicing contrast, although the difference in performance with the lengthened contrast /p+ b/ vs /p b/ was not significant. However, the fact that /p+ b/ was not significantly different from /p k/, but /p b/ was, also suggest a difference not reflected in the statistical inference: that detecting voicing without a concomitant duration difference is significantly harder for Kriol speakers than detecting a simple place difference. The performance of the Kriol listeners in Study 2 supports the conclusion drawn on the basis of the Wubuy participants' results: lack of native experience in a voicing contrast (for Kriol listeners: with fricatives only) leads to an inability to discriminate that contrast, even for L2 learners with extensive L2 exposure. Interestingly, however, in the case of the Kriol listeners, their experience *with* voicing contrasts in stops does not translate to

an ability to perceive this characteristic in fricatives. We return to this point in the conclusion.

#### 4.3. English results

Finally, a series of one-sample *t*-tests against chance performance indicated that—as is apparent from Figs. 2 and 3—the English listeners' discrimination of all six contrasts was significantly better than chance ( $p < .001$  in all cases). A final set of One-Way ANOVAs revealed there was no significant effect of contrast for either Study 1 ( $F(2,36) = 2.153, p = .131$ ) or Study 2 ( $F(2,36) = 1.003, p = .377$ ).

These results, unsurprisingly, provide evidence that the English listeners are well able to discriminate the obstruents in Study 1 and 2, as these straight-forwardly map onto their native categories. The fact that the discrimination accuracy for the Kriol-like /p+ b/ contrast is on par with the discrimination accuracy for the original English /p b/ contrast suggests that the listeners continued to pay attention to the VOT difference, and that the artificially lengthened constriction duration did not distract or aid their discrimination.

### 5. Conclusion

The results of the present studies show—unsurprisingly—that L1 background systematically shapes perception of L2 phonological contrasts that (1) do not align with L1 phoneme boundaries, or (2) differ drastically from the L1 phonemes in their phonetic realisation. Importantly, however, these studies also indicate that L2 phonemic learning may be near-impossible if a learner's L1 has not provided him/her with some familiarity with voicing (or with constriction duration-based) contrasts. They also indicate that familiarity with voicing in one phonemic domain (stops), does not necessarily translate to ability in another (fricatives), despite both of these categories being phonologically classified as obstruents, and thus in the category where (if anywhere) we expect voicing to be implemented phonemically. This result leads us to question the extent to which a phonological feature such as [ $\pm$ voice] can be said to be activated by native language input. This is despite the fact that the Wubuy and Kriol listeners are end-state learners, most likely; all have acquired English since school, and have to use it on a regular basis in certain domains (school, the shop, contact with government officials, etc).

In considering the difference in the performance of the English-speakers, as opposed to the Numbulwar residents, more generally, however, we must consider the nature of the task and the characteristics of the participants. The Wubuy and Kriol listeners likely faced significant levels of task fatigue and difficulty, compared to the English listeners, due to differences in schooling and lifestyle. Comparison of the Wubuy and Kriol speakers is less problematic, but compromised by the fact that they are distinct age-groups: the Wubuy listeners were much older.

### 6. Acknowledgements

We thank the Wubuy, Kriol and English listeners. We also thank the English speakers who provided the stimuli for the present studies, and Josh Clothier for assisting with the testing of the English control group. We thank Dr. Thomas Britz for Fig. 1. We gratefully acknowledge the Australian Research Council for generously supporting this research through DP130102624 'Learning to talk Whitefella Way'.

## 7. References

- [1] A. S. Abramson and L. Lisker, "Discriminability along the voicing continuum: Cross-language tests," in *Proceedings of the Sixth International Congress of Phonetic Sciences*, B. Hala, M. Romportl, P. Janota, Eds. Prague: Academia, 1970. pp. 569–73.
- [2] W. Strange, "The effects of training on the perception of synthetic speech sounds: voice onset time," Ph.D. dissertation, University of Minnesota, 1972.
- [3] D. Pisoni, R., Aslin, A., Perey, and B. Hennessy, "Some effects of laboratory training on identification and discrimination of voicing contrasts in stop consonants," *J Exp. Psych: Human Perception and Performance* vol. 8, pp. 297–314, 1982.
- [4] A. S. Abramson and L. Lisker, "Voice-timing perception in Spanish word-initial stops," *J Phon* vol. 1, pp. 1-8, 1973.
- [5] J. E. Flege, "Production and perception of English stops by native Spanish speakers," *J Phon* vol. 15, pp. 67-83, 1987.
- [6] C. T. Best, "The emergence of native-language phonological influences in infants: A perceptual assimilation model," in *The Development of Speech Perception: The transition from speech sounds to spoken words*, J. C. Goodman, H. Nusbaum, Eds. Cambridge, Massachusetts: MIT Press, 1994, pp. 167-224.
- [7] C. T. Best, "A direct-realist view of cross-language speech perception," in *Speech Perception and Linguistic Experience: Issues in cross-language research*, W. Strange, Ed. Timonium, MD: York Press, 1995, pp. 171-204.
- [8] C. T. Best, M. D. Tyler, "Nonnative and second-language speech perception: Commonalities and complementarities," in *Second Language Speech Learning: The role of language experience in speech perception and production*, J. Munro, O.-S. Bohn, Eds. Amsterdam: John Benjamins, 2007, pp. 13-34.
- [9] J. W. Harris, *Northern Territory Pidgins and the Origin of Kriol*, Series C, vol 89. Canberra: Pacific Linguistics, 1986.
- [10] J. Sandefur, *Kriol of North Australia: A language coming of age*. Darwin: Summer Institute of Linguistics, 1986.
- [11] J. Munro, "Roper River Aboriginal language features in Australian Kriol: Considering semantic features," in *Creoles, their Substrates, and Language Typology*, C. Lefebvre, Ed. Amsterdam: John Benjamins, 2011, pp. 461–87.
- [12] Australian Institute of Aboriginal and Torres Strait Islander Studies/Commonwealth of Australia. 2005. *The National Indigenous Languages Survey Report*. Canberra: Department of Communications, Information Technology and the Arts.
- [13] J. Hudson, *Grammatical and Semantic Aspects of Fitzroy Valley Kriol*, Work Papers of SIL-AAB A8. Darwin: Summer Institute of Linguistics, 1983.
- [14] B. Baker, R. L. Bundgaard-Nielsen, and S. Graetzer, "The obstruent inventory of Roper Kriol," *Australian Journal of Linguistics*, vol. 34, no. 3, pp. 307-344, 2014.
- [15] J. Heath, *Functional Grammar of Nunggubuyu*. Canberra: Canberra: Australian Institute of Aboriginal Studies, 1984.
- [16] J. Fletcher and A. Butcher, "Sound patterns of Australian languages," in *The Languages and Linguistics of Australia: A Comprehensive Guide*, H. Koch, R. Nordlinger, Eds. Berlin: Walter de Gruyter, 2014, pp. 91-138.
- [17] P. Boersma and D. Weenink, "Praat: doing phonetics by computer" [Computer program], Version 5.1.44. 2010.
- [18] J. D. Cohen, B. MacWhinney, M. Flatt, and J. Provost, "PsyScope: A new graphic interactive environment for designing psychology experiments," *Behavioral Research Methods, Instruments, and Computers*, vol. 25 no. 2, pp. 257-271, 1993.
- [19] J. Heath, *Ngandi Grammar, Texts, and Dictionary*. Australian Institute of Aboriginal Studies, 1978.